Coláiste na Tríonóide, Baile Átha Cliath
Trinity College Dublin
Ollscoil Átha Cliath | The University of Dublin

Information Theoretical Aspects of Complex Systems

Lecture 2.04

EEU45C09 / EEP55C09
Self Organising Technological Networks

CONNECT
Networks of the Future

Nicola Marchetti
nicola.marchetti@tcd.ie

Entropy

- Average information one gets when the system is observed

$$S[P] = \left\langle \log \frac{1}{p_i} \right\rangle_i = \sum_{i=1}^{n} p_i \log \frac{1}{p_i} \quad . \tag{1}$$

- Expected gain of information when we observe a system characterized by a probability distribution P over its possible states

- *Lack of knowledge of the system* (before the exact state of the system is observed)

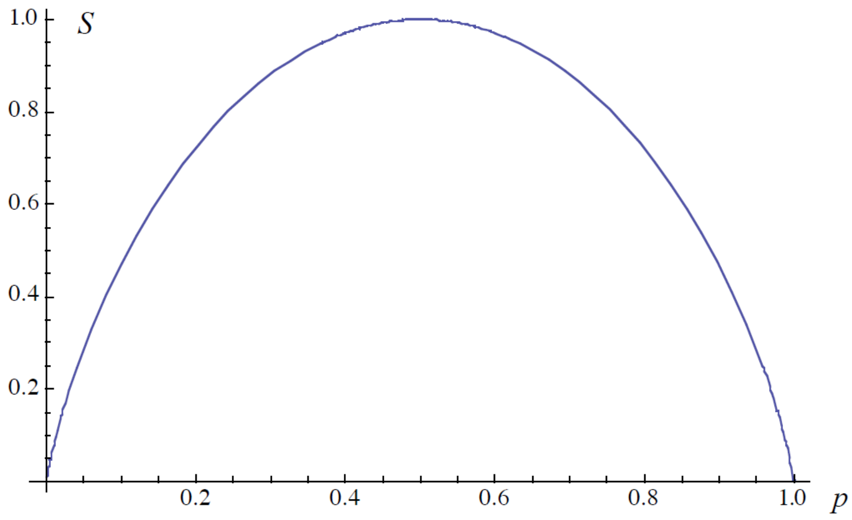
Example for n = 2 states

- Two possible states with probabilities p and 1-p

- Entropy S(p) – lack of knowledge – maximum when both states are equally probable, i.e., p = 1/2

- In such case we have no clue on which state we will find the system in, when observing it

n=2 case



What happens to the entropy when we know the system is in a certain state?

Entropy and coding – an example

- Stochastic process that generates random sequences of symbols a, b, c, d

- Suppose – to start with – that it is unknown with which probabilities symbols are generated

- Then the best guess is to assign probability 1/4 to each event, and that subsequent symbols are independent

- Then information gained in any possible observation of a single symbol is $\log_2(4) = 2$ bits. Therefore, the entropy $S = 2$ bits.

- Reasonable, as we can simply code our four symbols with binary codewords 00, 01, 10, 11

Entropy and coding – an example

- Suppose again symbols are generated independently of each other, but with probabilities p(a) = 1/2 , p(b) = 1/4 , p(c) = 1/8, p(d) = 1/8 and this is known a priory to the observer

- Since $I(p) = \log \frac{1}{p}$ when observing a we get 1 bit, but d would give us 3 bits

- Applying Eq. (1) we get that

$$S = 1/2 \; + \; 1/4 \cdot 2 \; + \; 1/8 \cdot 3 \; + \; 1/8 \cdot 3 \; = \; 7/4.$$

Entropy and coding – an example

- We find that the *entropy is now reduced, since we have some prior knowledge on the probabilities with which symbols occur*

- By making a better code than the trivial one mentioned before, we could use the codewords 0 for a, 10 for b, 110 for c, 111 for d

- In that case, the average codeword length decreases from S = 2 to S = 1.75 bits

- The trick here lies in the fact we used a codeword length (in bits) equal to the information gained if the corresponding symbol is observed

- Common symbols that carry little information should be given short codewords, and vice versa

Entropy as an additive quantity

- The *entropy of a system composed by independent parts equals the sum of the entropies of the parts*

- Special case : two subsystems characterized by

$$Q = \{q_i\}_{i=1}^{n} \quad \text{and} \quad R = \{r_j\}_{j=1}^{m}$$

- q_i and r_j represent probabilities for states i and j in the two subsystems, respectively

$$P = \{q_i r_j\}_{i=1,j=1}^{n,m}$$

Entropy as an additive quantity

$$S[P] = \sum_{i=1}^{n}\sum_{j=1}^{m} q_i r_j \log\frac{1}{q_i r_j} = \sum_{i=1}^{n}\sum_{j=1}^{m} q_i r_j \left[\log\frac{1}{q_i} + \log\frac{1}{r_j}\right] =$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} q_i r_j \log\frac{1}{q_i} + \sum_{i=1}^{n}\sum_{j=1}^{m} q_i r_j \log\frac{1}{r_j} =$$

$$= \sum_{i=1}^{n} q_i \log\frac{1}{q_i} + \sum_{j=1}^{m} r_j \log\frac{1}{r_j} = S[Q] + S[R]$$

Where we have used

$$\Sigma_i\, q_i = \Sigma_j\, r_j = 1$$

Kullback information

- When we make an observation, the exact microstate may not be revealed

- Based on the observation though we may replace our original (a priori) distribution $P^{(0)}$ with a new one P

- Information gained in the observation is called the Kullback information $K\left[P^{(0)};P\right]$ and is defined as

$$K\left[P^{(0)};P\right] = S^{(0)} - S = \sum_{i=1}^{n} p_i \log \frac{1}{p_i^{(0)}} - \sum_{i=1}^{n} p_i \log \frac{1}{p_i} = \sum_{i=1}^{n} p_i \log \frac{p_i}{p_i^{(0)}}$$

Kullback information and entropy

- $K\left[P^{(0)};P\right] \geq 0$ **(2)**

- Above follows directly from Gibbs inequality (*see Lecture 2.02*)

- We can use the fact that $K\left[P^{(0)};P\right] \geq 0$ to prove that the *entropy is a concave function*

- This means that, if P and Q are two probability distributions (both over n possible states), the entropy of any weighted average of P and Q is larger than the corresponding weighted average of their respective entropies:

$$S[a \cdot P + (1-a) \cdot Q] \geq a \cdot S[P] + (1-a) \cdot S[Q]$$

where $0 \leq a \leq 1$

Kullback information and entropy

$$S[a \cdot P + (1-a) \cdot Q] \geq a \cdot S[P] + (1-a) \cdot S[Q]$$

Proof.

$$S[a \cdot P + (1-a) \cdot Q] - \left(a \cdot S[P] + (1-a) \cdot S[Q]\right) =$$

$$= \sum_{i=1}^{n} (ap_i + (1-a)q_i) \log \frac{1}{ap_i + (1-a)q_i} - \sum_{i=1}^{n} \left(ap_i \log \frac{1}{p_i} + (1-a)q_i \log \frac{1}{q_i} \right) =$$

$$= a \sum_{i=1}^{n} p_i \log \frac{p_i}{ap_i + (1-a)q_i} + (1-a) \sum_{i=1}^{n} q_i \log \frac{q_i}{ap_i + (1-a)q_i} =$$

$$= aK[a \cdot P + (1-a) \cdot Q; P] + (1-a)K[a \cdot P + (1-a) \cdot Q; Q] \geq 0.$$

where the last step follows from Eq. (2).

Maximum Entropy Principle

- Even if we do not know exactly the state of a certain system, we may have some information on it

- For example, we could know the average energy or the number of particles

- Statistical mechanics is based on the idea that, with such limited information on the state of the system, we make an estimate of the probabilities for the possible microstates

- Alas, usually there are an infinite number of possible probability distributions consistent with the known properties of the system under study

Maximum Entropy Principle

- How should we then choose the probability distribution describing our system?
- Here it is reasonable to use the concept of entropy, since it can be interpreted as our lack of information on the system state
- When assigning a probability distribution for the system, we should not use one that represents more knowledge than what we already have
- Therefore, we choose – among the probability distributions that are consistent with the known system properties – the one maximizing the entropy

Maximum Entropy Principle

- The probability distribution can then be derived from a maximization problem with constraints

- The method assures that we do not include any more knowledge in the description of the system, than we already have

- This is the basic idea behind the Maximum Entropy Principle (MEP), also called the *principle of minimal bias*

Maximum Entropy Principle

- Therefore, if we assume a system for which we can measure certain *macroscopic* characteristics…

- And we assume that the system is made up of many *microscopic* elements, and that the system is free to vary among various states…

- And if we assume that with probability essentially equal to 1, the system will be observed in states with maximum entropy (*axiom*)…

- We will in this case be able to gain understanding of the system by applying the MEP and, using Lagrange multipliers, to derive formulas for certain aspects of the system

Maximum Entropy Principle

- Suppose we have a set of *macroscopic* measurable characteristics f_k, $k=1,\ldots,M$ (which we can think of as constraints on the system), which we assume are related to *microscopic* characteristics $f_i^{(k)}$ via

$$\sum_i p_i \cdot f_i^{(k)} = f_k$$

This means that we have M functions $f_i^{(k)}$, k = 1,…, M, of the microstates i, and that we know the expectation values of these, f_k. Such a function could, for example, give the energy of microstate i.

- As usual, we also have the constraints

$$p_i \geq 0$$
$$\sum_i p_i = 1$$

- We want to *maximise the entropy*, $\sum_i p_i \log(1/p_i)$ subject to the above constraints

Maximum Entropy Principle

- Using Lagrange multipliers λ_k (one for each equality constraint), we can show we find the general solution

$$p_i = \exp\left(-\lambda - \sum_k \lambda_k f_i^{(k)}\right)$$

- Since $\sum_i p_i = 1$ we can write

$$1 = \sum_i p_i = \sum_i \exp\left(-\lambda - \sum_k \lambda_k f_i^{(k)}\right)$$

$$1 = \sum_i \exp(-\lambda) \exp\left(-\sum_k \lambda_k f_i^{(k)}\right)$$

Maximum Entropy Principle

$$1 = e^{-\lambda} \sum_i \exp\left(-\sum_k \lambda_k f_i^{(k)}\right)$$

- And if we define

$$Z(\lambda_1, \ldots, \lambda_M) = \sum_i \exp\left(-\sum_k \lambda_k f_i^{(k)}\right)$$

we can then write

$$e^{\lambda} = Z \Rightarrow \lambda = \ln(Z)$$

Applying the MEP - Economics

- Suppose there is a fixed amount of money (M euros), and a fixed number of agents N in the economy

- Suppose that during each time step, each agent randomly selects another agent and transfers one euro to the selected agent

- An agent having no money does not go in debt

- What will be the long-term stable distribution of money?

- *Note*: this example depicts a not very realistic economy, as there is no growth, but only a redistribution of money

Applying the MEP – Economics

- We are interested in looking at the money distribution in the economy, so we are looking at the probabilities $\{p_i\}$ that an agent has the amount of money i, $i=0,\dots,M$

- We want to develop a model for the collection $\{p_i\}$

- If we let n_i be the number if agents who have i euros, we have two constraints

$$\sum_i n_i \cdot i = M \qquad \textbf{(3)}$$
$$\sum_i n_i = N$$

Applying the MEP – Economics

- Since we can write $p_i = n_i/N$ we can rewrite (3) as

$$\sum_i p_i \cdot i = \frac{M}{N}$$

$$\sum_i p_i = 1$$

- We now apply Lagrange multipliers

$$L = \sum_i p_i \ln(1/p_i) - \lambda\left[\sum_i p_i \cdot i - \frac{M}{N}\right] - \mu\left[\sum_i p_i - 1\right]$$

Applying the MEP - Economics

- Deriving L with respect to p_i we get

$$\frac{\partial L}{\partial p_i} = -[1 + \ln(p_i)] - \lambda i - \mu = 0$$

- We can solve this for p_i

$$\ln(p_i) = -\lambda i - (1 + \mu) \Rightarrow p_i \underset{\lambda_0 \equiv 1+\mu}{=} e^{-\lambda_0} e^{-\lambda i} \qquad \textbf{(4)}$$

Applying the MEP – Economics

- Putting in the constraints, we have

$$1 = \sum_i p_i = \sum_i e^{-\lambda_0} e^{-\lambda i} = e^{-\lambda_0} \sum_i e^{-\lambda i}$$
$$\frac{M}{N} = \sum_i p_i \cdot i = \sum_i e^{-\lambda_0} e^{-\lambda i} \cdot i = e^{-\lambda_0} \sum_i e^{-\lambda i} \cdot i \qquad \textbf{(5)}$$

- For large M, we can approximate as

$$\sum_{i=0}^{M} e^{-\lambda i} \approx \int_0^M e^{-\lambda x} dx \approx \frac{1}{\lambda}$$
$$\sum_{i=0}^{M} e^{-\lambda i} \cdot i \approx \int_0^M x e^{-\lambda x} dx \approx \frac{1}{\lambda^2} \qquad \textbf{(6)}$$

Check this.

Applying the MEP - Economics

- Substituting (6) into (5) we have (approximately)

$$e^{\lambda_0} = \frac{1}{\lambda}$$
$$e^{\lambda_0}\frac{M}{N} = \frac{1}{\lambda^2}$$

- From this, we get

$$\lambda = \frac{N}{M} = e^{-\lambda_0}$$

and thus, letting $T = M/N$, from (4) we get

$$p_i = e^{-\lambda_0} e^{-\lambda i} = \frac{1}{T} e^{-\frac{i}{T}}$$

Boltzmann-Gibbs distribution

By analogy, we can think of *T* (the average amount of money per agent) as the *temperature* - we have what is called a *Boltzmann economy*

Generalisation to a continuous state space

- Information theory also applies to a continuous state space

- Assume we have a probability density p($\mathbf{x}$) over this state space, where $\boldsymbol{x} = (x_1, \cdots, x_D)$ is a vector in a D-dimensional Euclidean space E

- This could – for example – mean that we do not know the position of a particle, but that we describe such position as a probability density over this space

- The probability to find the system in certain volume V is then

$$P(V) = \int_V d\mathbf{x}\, p(\mathbf{x})$$

Generalisation to a continuous state space

- The probability normalisation constraint requires

$$\int_{\mathbf{E}} d\mathbf{x}\, p(\mathbf{x}) = 1$$

- We can then define entropy as

$$S[p] = \int d\mathbf{x}\, p(\mathbf{x}) \ln \frac{1}{p(\mathbf{x})}$$

- And the Kullback information as

$$K\left[p^{(0)}; p\right] = \int d\mathbf{x}\, p(\mathbf{x}) \ln \frac{p(\mathbf{x})}{p^{(0)}(\mathbf{x})}$$

Ackowledgement

- Kristian Lindgren, "Information Theory for Complex Systems", pages 4-12